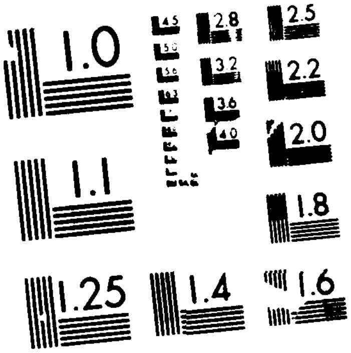
AD-A187 680 ANALYSIS OF LEARNING CURVE FITTING TECHNIQUES(U) AIR FORCE INST OF TECH WRIGHT-PATTERSON AFB OH SCHOOL OF SYSTEMS AND LOGISTICS C R AVINGER SEP 87 1/1

UNCLASSIFIED AFIT/GSM/LSQ/87S-3 F/G 5/3 NL


1.0
4.5
5.0
5.6
6.3
2.8
3.2
3.6
4.0
2.5
2.2
2.0
1.1
1.8
1.25
1.4
1.6

6

ANALYSIS OF LEARNING CURVE FITTING TECHNIQUES

THESIS

Charles R. Avinger
Captain, USAF

AFIT/GSM/LSQ/87S-3

A
N
D
U

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

2

AFIT/GSM/LSQ/87S-3

ANALYSIS OF LEARNING CURVE FITTING TECHNIQUES

THESIS

Charles R. Avinger
Captain, USAF

AFIT/GSM/LSQ/87S-3

Approved for public release; distribution unlimited.

The contents of the document are technically accurate, and no sensitive items, detrimental ideas, or deleterious information is contained therein. Furthermore, the views expressed in the document are those of the author and do not necessarily reflect the views of the School of Systems and Logistics, the Air University, the United States Air Force, or the Department of Defense.

AFIT/GSM/LSQ/87S-3

ANALYSIS OF LEARNING CURVE FITTING TECHNIQUES

THESIS

Presented to the Faculty of the School of Systems and Logistics
of the Air Force Institute of Technology
Air University
In Partial Fulfillment of the
Requirements for the Degree of
Master of Science in Cost Analysis

Charles R. Avinger
Captain, USAF

September 1987

Approved for public release; distribution unlimited.

Preface

This research studied the ordinary least-squares, weighted least-squares, median slope and mean slope techniques for fitting data with learning curve theory. The basic idea was to find the best technique to use for equal and unequal lot data with normal, triangle and Cauchy error term distributions. Along the way, the Cauchy distribution proved to be too unwieldy and was dropped.

The analysis showed that no technique was best in any particular situation, each technique had tradeoffs among t h estimates of the parameters (first unit cost and learning curve coefficient), the dispersion of the data points around the fitted line and the range and bias of future predictions. However, the data should improve your understanding of these techniques and their limitation when you use them in the future.

I wish to thank my thesis advisor, Roland D. Kankey, for his tireless efforts in making this the best research that it could be. I also wish to thank my wife, Michelle, for putting up with all my fits when I had 50 computer runs spread throughout the house and couldn't find the one I wanted, and for her efforts in helping me get the thesis finished in a timely fashion.

Charles R. Avinger

Table of Contents

List of Figures

List of Tables

AFIT/GSM/LSQ/87S-3

Abstract

This thesis was basic research on fitting techniques for learning curve. The unit formulation of the learning curve theory was fit using two parametric fitting techniques, ordinary least-squares and weighted least-square, and two nonparametric fitting techniques that were called median slope and mean slope. Comparisons were made between the techniques in four data cases, equal lot sizes with normally distributed error terms, unequal lot sizes with normally distributed error terms, equal lot sizes with triangularly distributed error terms and unequal lot sizes with triangularly distributed error terms. The possibility that error terms could be distributed based on the Cauchy distribution was tried and rejected.

The data for these four cases was simulated in the SAS System and based on a first unit cost of 25,000 and an 80% learning curve slope. The normal error terms used a standard distribution of .12 (in logarithmic form) and the triangle error terms used a range of -.36 to .36 (in logarithmic form) to include all error terms. The learning curve heuristic was used to determine lot plot points.

The fitting techniques were compared on their estimation of the parameters (first unit cost and the learning curve coefficient), the dispersion of the data points to the fitted line and the predictions of future costs.

Analysis showed that in cases of equal lot sizes the ordinary least-squares technique has more bias than the nonparametric techniques, but still is a good estimator. In cases of unequal lot sizes the weighted least-squares and the ordinary least-squares techniques performed well. Ordinary least-squares estimated the parameters with the least bias but had average predictions errors that increased with the unit number. Weighted least-squares estimated the parameters with the most bias and had predictions errors that, on average, started high and turned low as the unit number got larger. The nonparametric techniques did a better job of predicting future costs, in that the mean predictions were closer to the population costs. However, these techniques had wider ranges of predictions.

Because this was basic research, there are many areas for further research. These areas include the study of nonlinear regression, weights other than lot size for the weighted least-squares technique and statistical testing for bias based on increased production runs of 500 or 1000 fitted lines.

ANALYSIS OF LEARNING CURVE FITTING TECHNIQUES

I. Introduction

General Issue

Department of Defense and Air Force leaders are concerned with the management and control of cost to procure weapons systems. Accurate cost forecasting or prediction is a vital element of cost management and control. Proper estimation of production costs, the most significant portion of acquisition costs, is critical if DOD decision makers are to properly choose between the various alternate force weapon systems. As Gansler (6:3) and Bryan and Clark (3:105) indicate, military acquisition programs in the recent past have all too often encountered uncomfortable levels of cost growth. Since errors in cost estimates are contributors to this cost growth, accurate cost estimation is vital if cost growth is to be minimized. The identification and analysis of alternate cost estimating techniques, and the comparison among and between existing techniques, contributes to the development of the profession.

Specific Issue

Weapon system production costs fall into two categories. There are recurring costs, such as labor and material, and non-recurring costs, such as factory setup. Historically, recurring costs are the largest part of the costs to produce weapon systems. So, estimation of these recurring costs is very important. Recurring production costs are estimated with the use of the learning curve theory. This ubiquitous technique has achieved the status of a standard accepted cost analysis tool. The learning curve theory, sometimes referred to as the cost improvement curve theory, states that costs decrease in a regular pattern as more units are produced (11:16). For example, the unit formulation of the learning curve states that costs to produce a unit decrease by a constant percentage as the cumulative number of units produced doubles (1; 19). If the cost to produce the third unit was $10 and costs reduce by 10 percent between doubles, the sixth unit would cost $9 and the twelfth unit would cost $8.10.

The most common way to estimate the learning curve function from actual data uses a least-squares fitting technique. Most who use this technique then assume that errors follow a normal (or Gaussian) distribution. The normal probability distribution is assumed to account for the random fluctuation in costs. This random fluctuation is commonly referred to as the random error term. While this technique is very popular, there are other techniques of fitting a mathematical function to the data and other probability distributions that could be used. This thesis compares results of various curve fitting techniques given different error term probability distributions, determines the situations where each technique works best and compares the advantages and disadvantages of each.

Investigative Questions

This thesis covers three main areas. First, identification of the parametric techniques currently used to determine learning curve functions and identification of nonparametric techniques that could be used to determine learning curve functions. Second, identification of probability distributions for the random error terms. This identification includes the currently used probability distributions and other potential distributions. Third, evaluation of the fitting and forecasting performance of each technique under different conditions. Results of the selected curve fitting techniques are compared over various combinations of error term probability distributions and lot sizes of production units.

Background

The learning curve theory was first published in 1936 by Wright (19). Wright advanced the theory that cumulative average costs to produce aircraft decreased by a constant percentage as the number of aircraft produced doubled. He also used his theory to review the cost of producing automobiles. This theory has been applied to many production items since 1936. Learning curves are used extensively throughout DOD for cost estimating, contracting, planning and scheduling.

There are two formulations that are commonly used for learning curve work (2; 1). First, the unit formulation, which states that the cost to produce each unit declines by a constant percentage as the number of units produced are doubled. Second, the cumulative average formulation, which states that the average cost of all units produced declines by a constant percentage as the number of units produced doubles. The unit formulation was selected for use in this research because it is favored by theoreticians and practioners; and is most tractable.

The equation for the unit formulation of the learning curve is:

$$Y = A * X^B \qquad (1)$$

where

Y = the cost of each unit
A = the cost of the first unit
X = the sequential unit number in the production. For example X=101 says that the formula will be used to compute the cost of unit number 101
B = a constant related to the rate of cost improvement (2: Chap 7,8-9), this number will be called the learning curve coefficient in this thesis

The unit formulation of the learning curve is based on the premise that a constant percentage change in the cost of a unit will drive a constant percentage change in the unit cost. This means that formula (1) produces a straight line on log-log graph paper where equal distances between numbers represents equal percentage changes (1:10-12). Formula (1) is thus often said to be a log-linear function (see Figures 1 and 2 on page 6). Because of this logarithmic property, formula (1) can also be written as:

$$Y' = A' + B * X' \qquad (2)$$

where

$Y' = \ln (Y)$
$A' = \ln (A)$
$X' = \ln (X)$
ln = the natural logarithm function

and is called the linear form of the unit formulation. This second formula is derived from formula (1) by taking the natural logarithm of each side of the equation. Formula (2) can be plotted on standard graph paper and will produce a straight line. One term peculiar to the theory is the idea of a learning curve slope. A ten percent rate of learning (or reduction in cost) between doubled quantities is equivalent to a "learning curve slope" of ninety percent, a twenty percent rate of learning is equivalent to an eighty percent slope. For the purposes of this research it is sufficient to note that there is a direct relationship between the B parameter, the learning curve slope and the learning rate. Given any one, the others can be uniquely identified. For a more detailed discussion of the learning curve theory see Kankey's article, "Learning Curves: A

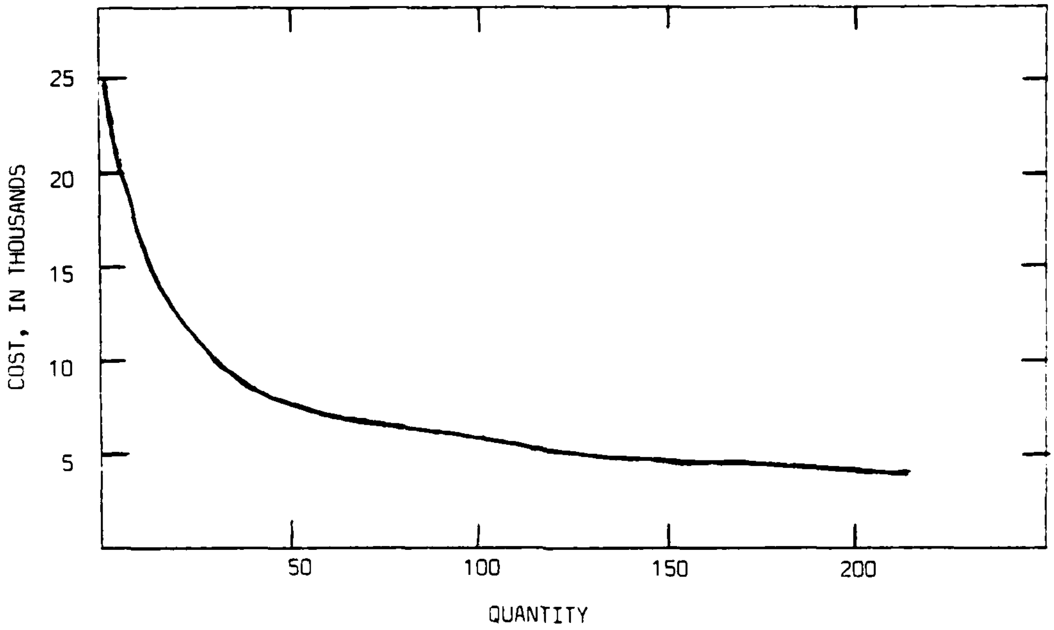


Figure 1. Learning Curve
on Standard Graph Paper

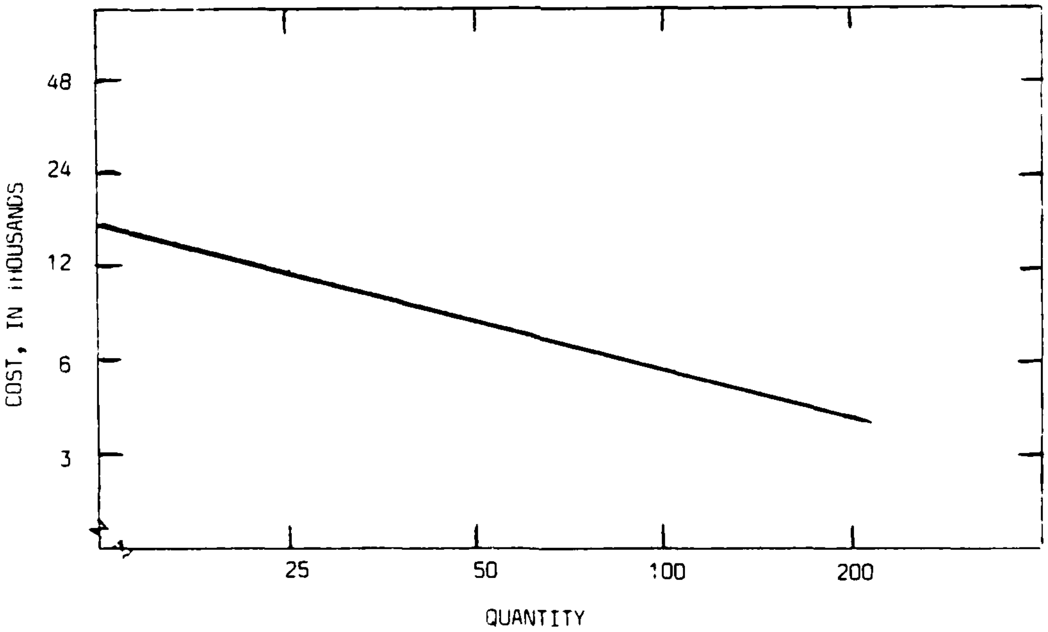


Figure 2. Learning Curve
on Log-Log Graph Paper

Review" (11:16-19); Wright's article, "Factors Affecting the Cost of Airplanes" (13:122-128); Volume 1, "AFSC Cost Estimating Handbook" (2:Chap 7); or Hayes and Wheelwright's chapter, "The Experience Curve" in their book Restoring Our Competitive Edge: Competing Through Manufacturing (7:229-274).

Limitations and Assumptions of the Study

Limitations.

1. Only the unit formulation is studied in this thesis.

2. Only one slope for the population is considered. All simulation is based on one learning curve slope.

3. Only the case of a standard multiplicative error term is considered (see 2. below under Learning Curve Assumptions).

4. Lot plot points are determined using the learning curve heuristic (see the Equal Lot Data section of Chapter 3 for an explanation of this heuristic).

Learning Curve Assumptions.

1. The population follows the unit formulation of the learning curve theory.

2. The unit formulation of the learning curve is specified in standard form by:

$$Y = A * X^B * E \qquad (3)$$

where

Y = the cost of the xth unit
A = the cost of the first unit
B = a constant related to the rate of cost improvement, this number will be called the learning curve coefficient in this thesis.
X = the unit number
E = the multiplicative error term.

This is a multiplicative model which is log-linear. Therefore, any probability distribution of the error term will be specified in the transformed state, that is to say in the linear form.

3. The unit formulation of the learning curve is specified in linear form by:

$$Y' = A' + B * X' + e \qquad (4)$$

where

Y' = the natural logarithm of the cost of the xth unit
A' = the natural logarithm of the cost of the first unit
B = the same as in formula (3)
X' = the natural logarithm of the unit number
e = the random error term = $\ln(E)$ from formula (3)
A' and B' are the parameters to be estimated

Model Assumptions.

1. The expected value of the cost of unit x can be written:

$$E(Y'_x) = A' + B * X' \qquad (5)$$

where

$E(Y'_x)$ = the expected value of y given x
the other variables are defined the same as in equation (4)

2. The random error terms, e, are independent. That is to say, the value of the error term at a given x does not depend upon the value of the error term at any other x.

3. The random error terms, e, come from populations each having a mean of zero and a constant variance. This implies that any probability distribution of the random error term applies to the unit learning curve formula in the linear form, formula (4). For instance, if the normal probability distribution is specified in the linear form, the distribution is called log-normal in the standard form of the model. It should be noted that a log-normal distribution is very different from a normal distribution.

Definitions

Random error term - the *difference between* the actual value of an item and the expected value of that item. The random error can be viewed as the portion of the cost of the xth unit not specified by formula (3). That is to say, the effect of any variable, other than the unit number, on the cost of a unit contributes to the error term. For example, the error term associated with your height would be the difference between your height and the average height of all people. If your height was 73" and the average height of all people was 71", the error term would be the 2". When dealing with the unit learning curve, the random error term would be the difference between the actual cost of the xth unit and the expected cost of the xth unit as specified by formula (5).

Nonparametric statistics - statistics that require few assumptions about underlying populations, most notably the assumption about the normal distribution of the population (8:1). Nonparametric statistical tests do not require a normal population assumption and are generally easier to use and understand according to Hollander and Wolfe (8:1). They go on to state that, based on theoretical investigations, these nonparametric statistics do not usually forgo much in comparison to parametric statistics. Conover states:

> A statistical method is nonparametric if it satisfies at least one of the following criteria:
> 1. The method may be used on data with a nominal scale of measurement.
> 2. The method may be used on data with an ordinal scale of measurement.
> 3. The method may be used on data with an interval or ratio scale of measurement, where the distribution function of the random variable producing the data is either unspecified or specified except for an infinite number of unknown parameters. (5:94)

Conclusion

The importance to DOD of accurate estimation of weapons systems cost cannot be overstated. The recurring production costs of these weapons systems are a major part of the total life cycle cost of the weapon system. This thesis analyzes several aspects of the method almost universally used to predict the recurring cost, learning curve theory. The unit formulation of the learning curve theory is analyzed with both parametric and nonparametric techniques.

II. Literature Review

Two literature reviews were performed, one to identify fitting techniques for the learning curve and the other to identify potential probability distributions for the error term. An attempt to find previous studies comparing these fitting techniques proved fruitless, no previous studies were located.

Fitting Techniques

A literature review was performed to identify the fitting methods or techniques used to determine the mathematical formula of a learning curve given existing data. The review identified the current techniques used to estimate the learning curve parameters and some additional techniques that could be used. Advantages and disadvantages of each technique were also identified. All fitting techniques identified fit linear data so the transformation from log-linear to linear is necessary before the intercept (A) and the slope (B) parameters can be estimated. Each technique estimates the value of the a and b statistics (estimates of the transformed parameters A' and B') for formula (2) based on existing data. Traditionally, population parameters are denoted with Greek letters, e.g. alpha and beta, but I will use capital letters, e.g. A and B. Statistics, or estimates of these parameters are Roman, e.g. a and b.

Current Techniques. Currently, there are two techniques that are commonly used to estimate the learning curve formula, ordinary least-squares and weighted least-squares (2:Chap 7). Both of these techniques are among the parametric techniques commonly referred to as regression techniques.

Ordinary least-squares is the most common technique (2:Chap 7). Least-squares is a parametric technique that has some underlying assumptions. These assumptions were discussed as model assumptions in the Introduction chapter. The least-squares technique determines the a and b statistics for formula (2). This determination is done by summing the squared differences between each lot's average actual cost per unit and the average cost of each lot estimated by the improvement curve formula. For a more detailed explanation of the ordinary least-squares technique, see Neter, et. al., Applied Linear Regression Techniques (15:23-52). Random errors are assumed to be normally distributed when using ordinary least-squares, according to Johnston in Econometric Methods (10:168-171,181). The main advantages of the least-squares technique is the ease of use, the ability to get quick answers, and the technique's ability to still give good answers when some of the assumptions about the technique are violated a small amount. The main disadvantages are that the technique is not theoretically correct when the lots have different weights and that the technique requires a specified probability distribution for the error terms before any testing or interval estimating (15:48). Lots have equal weights when the same number of units are in each lot. When lots have different numbers of units in each lot, the lots have different weights. For the specific learning curve case, the normal probability distribution assumption causes the error term to be normally distributed in the log-linear or transformed state (as in formula (4)).

Weighted least-squares is used when the data has unequal production lot sizes. For example, weighted least-squares would be used if the data contained the average cost of each lot and the first lot contained 20 aircraft, the second lot contained 50, and the next seven lots contained 100 each. Weighted least-squares would give less emphasis to the first two lots since they are small and more emphasis to the last seven lots. The weighted least-squares method of dete·mining the statistics a and b would weight each average cost based on the number of units in the lot (15:167-172). The main reason behind using weighted least-squares rather than ordinary least-squares is because the expected variance of lot costs is not the same (15:168). This use of weighting is an application of the Central Limit Theorem, which states:

> If $Y_1, \ldots, Y_n$ are independent random observations from a population with probability function $f(\overline{Y})$ for which [the variance] sigma-squared is finite, the sample mean Y:
>
> $\overline{Y}$ = the sum from i = 1 to n of Y_i/n
>
> is approximately normally distributed when the sample size n is reasonably large, with mean E(Y) and variance (sigma-squared)/n (15:6).

According to Murphy, it is easily shown that the variances of each lot are unequal, even if the random error terms have the same variance (14). From the above theorem, if the random error for an aircraft were 20000, then the random error for the first lot would be 1000 (20000/20), the random error for the second lot would be 400 (20000/50), and the random error for the third lot would be 200 (20000/100). Murphy explains that the variance of the average cost of each lot equals the variance of the random error term divided by the lot size. When lot sizes are different, the variances must be different if the random error terms have equal variances as the assumptions require. The main advantages of the weighted least-squares best fit technique are the same as the advantages of the least-squares best fit technique. The weighted technique reportedly gives more correct parameter estimates when lots have different weights, according to Neter, et. al. (15:167-172). The weighted technique gives the same answer as the ordinary least-squares technique when the lots have equal weights. The main disadvantage is the assumption of an error term with a normal probability distribution. There is also the question of whether using lot sizes for weights overstates/overweights the large lots when the data is transformed into logarithms. For example, a lot composed of units 11 through 20 includes a doubling (100% increase) while a lot composed of units 150 through 200 has only a 30% increase. Should the larger lot get five times the weight of the smaller lot? This question is deferred to follow-on research.

Other Techniques. The literature review disclosed two methods of determining the parameters for formula (2) when nonparametric techniques are used. The first technique is identified by Hollander and Wolfe and by Conover (8:200-206; 4:263-271) and will be called the median slope technique. The second technique is identified by Hollander and Wolfe (8:206-208) and will be called the mean slope technique. Both of these techniques were tested in this thesis and both are nonparametric because no probability distribution is specified. This lack of a probability distribution meets the third condition Conover defines for a nonparametric statistical method (see page 10). Additional techniques that were not used in this thesis were reviewed.

The main advantages of nonparametric techniques, according to Conover (5:3), are that the probability distribution of the error term does not have to be known or assumed, a "simple and unsophisticated" model is used and the application of the technique requires less math because the model is simple. The main disadvantage is that, since the probability distribution is not known, estimates of equation parameters and estimates of future costs will have a much wider confidence interval. No other specific advantages or disadvantages can be attached to the median slope or the mean slope technique.

The median slope technique suggests that the points be put in order from lowest x to highest x, for convenience of analysis, and then requires that the slope between each pair of points be computed (8:205), reference Figure 3. (Note: for this thesis these points are expressed in logarithms.)

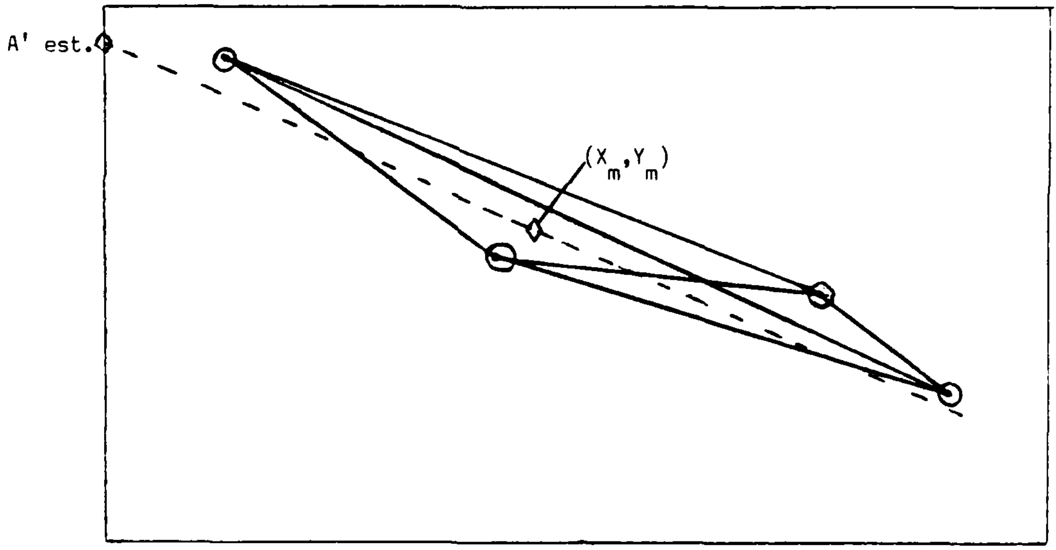


Figure 3. Illustration of Median Slope Technique

These slopes are then rank ordered and the median slope taken as the estimate of the B parameter from formulas (1) and (2). The A' parameter is estimated by using the median x and median y values (4:266-267), as indicated by the dotted line in Figure 3. When there are an even number of points, the average of the two middle slopes is used and the average x and average y associated with the two middle points are used (8:205; 4:266-267).

The mean slope technique also uses the slopes between each pair of points (8:205-208). A simple average slope is computed by summing the slopes and then dividing by the number of slopes. The mean slope technique uses the same estimation approach for the A' parameter as the median slope technique (8:206-208).

The four fitting techniques discussed will be used to estimate the unit formulation parameters of the learning curve in this thesis. A technique that was identified, but was not used is that of nonlinear regression. Nonlinear regression, according to Neter, et. al., is an iterative process that can require a great deal of time to solve (15:466). Evaluation of nonlinear regression for learning curve fitting would be an interesting area for further study, but was not included in this research.

Probability Distributions

The literature review allowed selection of three probability distributions for the error term. The normal probability distribution is currently assumed for the distribution of the error term. The review identified two other probability distributions that could be used, the triangle distribution and the Cauchy distribution. These two additional distributions were identified on the belief that the normal distribution was reasonable, or people would not be using it, and any other distribution identified should be similar to the normal distribution.

Normal Distribution. The normal probability distribution is usually assumed for the error terms. The normal distribution is smooth, symmetrical, continuous and bell shaped. It requires that, as the sample size gets very large, a specific percent of the random errors fall within one standard deviation (68.26%), within two standard deviations (95.44%), and within three standard deviations (99.74%) (13:201-202; 15:517).

Triangle Distribution. The triangle probability distribution can be symmetrical or asymmetrical. This distribution is smooth with a peak at the most likely point, continuous with a minimum and maximum possible value and shaped like a triangle. The shape of the distribution can be specified by the distribution parameters; lowest point, most likely point (where the distribution will peak), and the highest point (16:269). When the distance between the lowest point and the most likely point equals the distance between the highest point and the most likely point, the distribution will be symmetrical, otherwise the distribution is skewed. When using the triangle distribution with the learning curve only the symmetrical distribution will be used.

Cauchy Distribution. The Cauchy probability distribution is related to Student's t distribution, as both distributions are forms of a Pearson Type VII distribution (9:13, 154). The family of Pearson distributions are all related to the normal distribution, according to Johnson and Kotz (9:9-15). The Cauchy distribution is symmetrical, continuous, smooth, and shaped similar to the normal distribution. The Cauchy distribution has "longer and flatter tails (9:154)" than the normal distribution. (Note that flatter tails are often also called fatter tails.) Johnson and Kotz specify that the Cauchy distribution, with location and scale parameters equal to 0 and 1, respectively, "is the Student's t distribution with 1 degree of freedom (9:156-157)."

The literature review for techniques to fit learning curve data yielded four techniques that were used in this thesis: ordinary least-squares, median slope, mean slope and weighted least-squares. The literature review for the probability distributions identified three error term distributions that were used for this thesis: normal, triangle and Cauchy.

III. Methodology

The analysis of fitting techniques was a three step process. First, costs for production run lots were simulated using the SAS software system for data analysis, using known model parameters A and B and a fixed error term distribution. Second, the various techniques of fitting the learning curve were used to estimate the model parameters. Third, analysis of the statistics, dispersion and forecasting ability were performed. The various techniques were compared for the different lot data and for the different error term distributions.

The true relationship was selected to be an 80% slope learning curve with a first unit cost (A) of 25,000. A learning curve slope of 80% was selected as a typical slope for airframe assembly operations. An 80% learning curve slope yields a B value of -.321928095. The general objective was to see how well each of the fitting techniques estimated the population or true relationship (A and B) and estimated future costs under each of six cases. An ideal estimating technique would, on average, estimate A and B correctly, have a small dispersion of values and forecast with a tight, symmetrical distribution around the true value. The six cases considered in this study are based upon combinations of lot sizes, and error term distributions. These six cases are:

- Case I: Equal lot sizes, normal errors
- Case II: Unequal lot sizes, normal errors
- Case III: Equal lot sizes, triangle errors
- Case IV: Unequal lot sizes, triangle errors
- Case V: Equal lot sizes, Cauchy errors
- Case VI: Unequal lot sizes, Cauchy errors

Simulation of Data

The SAS System was used to simulate the data for lots from a production run. First, learning curve data was generated for each production run. Second, the data in each production run was separated into lot data, with equal size lots. Third, unequal lot sizes were identified and each production run was broken into unequal lots. These data creation steps involved the use of many SAS functions.

Production Runs. The simulation of production runs involved the use of three SAS random error generation functions, each based on a different probability distribution. A program was written to generate costs from a production run for units 1 through 210. This number of units was considered to be long enough to allow for valid comparisons, yet not overstrain the capability of the computer system for both space and processing time. The data was generated in the log-linear state and then transformed to the standard state. The data was generated with the formula:

Y' = A' + B * X' + e (6)

where

the variables are defined in formula (4) with
A' = ln(25000) = 10.12663110
B = ln(.8) / ln([illegible] = -0.321928095
X' = ln(units 1 through 210)
Y' = ln(cost)
except where the error term is
e = a random error term times a sizing value

The SAS function for random errors from a normal distribution is RANNOR(seed) (16:267-268). *The SAS User's Guide: Basics* states that this function will generate a value from a normal distribution with a mean of zero and a standard deviation of one. The seed is a number used to start the random number generation process. For the normally distributed random error in formula (6), the error formula was (16:267-268):

e = RANNOR(seed) * sigma (7)

where

seed = 1446
sigma = the standard deviation

For the triangular distribution random error in formula (6), the error formula was (16:269):

e = (BT-AT) * RANTRI(seed,(CT-AT)/(BT-AT)) + AT (8)

where

seed = 1346
BT = the highest triangle distribution value
AT = the lowest triangle distribution value
CT = the most likely triangle value = 0

For the Cauchy distributed random error in formula (6), the error formula was (16:265):

$$e = \text{RANCAU(seed)} * \text{sigma} \qquad (9)$$

where

seed = 446
sigma = a scale parameter similar to the normal distributions standard deviation

The values for Y' and X' were transformed to regular terms with the following equations:

$$Y = \text{EXP}\ (Y') \qquad (10)$$

$$X = \text{EXP}\ (X') \qquad (11)$$

where

EXP (Y') = $e^{Y'}$ i.e. Y' equals the natural logarithm of Y.
Y = the simulated cost
X = the unit number associated with the simulated cost

Data was simulated for 100 production runs using each of the three error term distributions.

Equal Lot Data. The data for each of the 100 production runs for each of the three error term distributions were turned into lot data with each lot size equal to 30 units. In addition, the lot plot point and mean cost of the units in each lot were determined. The lot plot point was determined based on the standard learning curve heuristic where the plot point is equal to half of the lot size plus the value of the last unit before the lot started. When the lot size is 30 and the first unit of the lot was unit 121, the lot plot point is 135 (30 / 2 + 120). When the first lot has 10 or more units the lot plot point is computed by the number of units in the lot divided by three.

Unequal Lot Data. The data for each of the 100 production runs for each of the three error term distributions was also used to get unequal lot data. The sizes of the first six lots were determined using the SAS function, RANUNI(seed) in the following equation (16:236-238,269):

$$\text{lot size} = \text{scaling factor} * \text{RANUNI(seed)} \quad (12)$$

where

scaling factor = a number based on the proposed lot size
seed = 1515
RANUNI(seed) = generates a uniformly distributed value between 0 and 1

Lot sizes were in the ranges shown in Table 1.

Table 1. Range of Lot Sizes

Lot Number	*Smallest* Lot Size	*Largest* Lot Size
1	2	10
2	15	25
3	20	30
4	25	35
5	30	40
6	40	50
7	20	78

Lot seven was the remainder of the units needed to bring the total number of units to 210. Lot plot points and the mean lot cost were determined the same way as for equal lot data.

Estimation of Learning Curve Formulas

For each production run a learning curve formula was estimated using the fitting techniques discussed in the literature review. These fitting techniques are ordinary least-squares, weighted least-squares, median slope and mean slope.

Ordinary Least-Squares (15). The ordinary least-squares technique was run on the SAS system. The data was converted to linear form with the following transformations:

$$y = \ln (Y) \qquad (13)$$

$$x = \ln (X) \qquad (14)$$

where

Y = the average cost of a production lot
X = the unit number of the median unit in the production lot
ln (variable) = the natural logarithm function

The ordinary least-squares procedure that was run is called PROC REG (17:658). The statements required were:

```
        PROC REG;
          MODEL y = x;
```

where the statements are defined in the SAS User's Guide: Statistics (17:658-659).

Weighted Least-Squares (15). The weighted least-squares technique was also run on the SAS system. This technique was the same as the technique for ordinary least-squares except the x transformation was:

$$x = m * (\ln (X)) \qquad (15)$$

where

the variables are defined for formula (14) and
m = the number of units in the lot

The weighted least-squares technique was only run on production data with different lot sizes because, as Neter states, the estimated parameters from weighted least-squares equal the estimated parameters from ordinary least-squares when the lot sizes are equal (15:167).

Median Slope (8:200-208; 4:263-271). Since no program existed, a program was written on the SAS system. First, the program transformed the data as was done in formulas (13) and (14). Second, the program calculated the slope between each lot of a production run from first lot to last lot, for the seven lots there were 21 slopes. Third, the slopes were rank ordered and the median slope picked. Fourth, the program determined the median lot and then determined the value of the a statistic from the following formula:

$$a = y_m - b_m * x_m \qquad (16)$$

where

y_m = the average cost from the median lot in logs

x_m = the x value from the median lot in logs

b_m = the median slope, estimate of the parameter B

a = the first unit cost statistic, estimate of the parameter A'

Mean Slope (8:206-203). Again, no program existed so a program was written on the SAS system. The program used the slopes previously determined and computed their mean. Then the value of the a statistic was determined fr.m formula (16) with one exception. The value of b_m was the mean of the slopes.

Analysis of the Curve Fitting Techniques

The analysis of the formulas generated using the simulated data was done in three areas. First, the distribution of the statistics about the population or true parameter value were compared for each case. Second, the fitting techniques for each case were compared based on two measures of dispersion used by forecasters, the mean absolute deviation (MAD) and the mean squared deviation (MSD). Third, the dispersion of the predicted cost of future units was compared for each of the fitting techniques within each case. Separate comparisons of each of the fitting techniques were made for each case.

Distribution of the Statistics. The distributions of the a and b statistics were reviewed. The 100 parameter estimates for each A and each B were displayed using Tukey's box-and-whisker type plots (18:39-42). These a and b statistics were from one series of 100 production runs for each case. Then the plots were compared. For instance, the box-and-whisker plots for the A parameter estimates from all fitting techniques for the equal lot, normal error term case were compared.

The Tukey box-and-whisker plot identifies the end points, the first (25%) and third (75%) quarter percentile points and the median point of a distribution. The style of the plot, as Tukey explains in Exploratory Data Analysis (18:32,39-41), is to draw a box from the first to the third quarter percentile points with a bar at the median and to add a separate line, or "whisker," to connect each end point to the box. This plot was modified by addition of the 5%, 95% and mean points. The plot and the data points were generated using the SAS procedure with plotting option,

PROC UNIVARIATE PLOT (16:1182,1187-1188)

The plots were then hand drawn from the data provided by the SAS procedure. Since converting these plots to publication quality is a very time consuming process, only a few sample plots are included in this thesis.

Dispersion of the Fitting Techniques. The learning curve formula represents the theoretical mean cost for each lot data point in the least-squares techniques and a similar idea in the other techniques. Variability of the data (13:29-30) about the estimated line is measured through use of the mean absolute deviation (MAD) and mean squared deviation (MSD). These variability measures plus many other measures can be used to reflect the accuracy of forecasts according to Kankey and Thompson (12:1-2). The MAD and the MSD were used to measure the raw error for each production run between the actual cost of each lot and the predicted cost of each lot based on the fitted formula. These measures were chosen because they are the most often used (12:3-4).

Mean Absolute Deviation. The MAD was computed for each fitting technique and each production run. Each individual deviation was computed with the formula:

$$D = Y - \hat{Y} \qquad (17)$$

where

Y = the actual average lot cost

$\hat{Y}$ = the average lot cost estimated

Then the absolute values of the deviations for a production run were summed. Finally, the sum of the absolute deviations was divided by the number of deviations. The average MAD for each of the different fitting techniques was compared for each case. Generally, techniques that have a lower MAD for the existing data are felt to be more likely to have lower absolute deviations when estimating.

Mean Squared Deviation. The MSD was computed the same as the MAD, except the deviation amounts were squared before the sum was computed. As with the MAD, the average MSD of each of the different fitting techniques were compared by case.

Dispersion of Predicted Values. The key ingredient in fitting learning curve data is to be able to predict future costs with accuracy. Costs of units 225 and 800 were predicted from each learning curve formula. The dispersion of the predicted values about the theoretical values was reviewed for each case.

Analysis of the Techniques. A comparison of the curve fitting techniques was made in three areas. Dispersion of the statistics, dispersion of the lot plot points from the fitted equation, and dispersion of the predicted values about the population or true value. These comparisons were used to compare the fitting techniques in each case reviewed.

IV. Findings

There were a number of interesting results. First, scaling of error term distributions was accomplished. Second, analysis of the fit provided by each technique was addressed. Third, the fitting techniques were analyzed for each case, with emphasis on advantages and disadvantages of each.

Error Term Distributions

Selection of Scale Values. An early step in the simulation of data was to determine the scaling for each error term distribution, i.e. standard deviation for the normal error term distribution; the scale parameter for the Cauchy error term distribution; and the highest, most likely and lowest points for the triangle error term distribution. The idea was to create data with a reasonable amount of deviation, a deviation that was also comparable with deviations in other cases.

To begin, the error term scaling parameters were selected. The determination of the standard deviation for the normal error term distribution was done through an iterative process. The first error level considered was 5% of the first unit cost. This resulted in a standard deviation of 1250 (25000*.05). This standard deviation would result in 99.74% (15:517) of all first unit costs falling within three standard deviations or between 0.0000128 and 4.88×10^{13} with a mean of 25,000. This amount of deviation was too extreme, costs would not be expected to vary this much. The range was computed using the formula:

$$EXP(Y' \pm (3 * e)) \qquad (18)$$

where

EXP( ) was defined in formulas (10) and (11)
$Y' = \ln(25{,}000) = 10.12663110$
$e = \ln(1{,}250) = 7.130898830$

The second standard deviation considered was 5% of the first unit cost in logarithmic form. This resulted in a standard deviation of .506331555 (10.1266311*.05) in logarithmic form. This standard deviation would again result in 99.74% (15:517) of all first unit costs falling within three standard deviations, or between 5,473 and 114,191 with a mean of 25,000, based on formula (18) with e = .506331555. The range was also considered too extreme. Cost of the first unit would not vary this greatly in most cases.

The third standard deviation considered was .12 in logarithmic form. This was selected subjectively based on reasonable magnitudes and resulted in 99.74% (15:517) of all first unit costs falling within three standard deviations, or between 17,442 and 35,833 with a mean of 25,000, based on formula (18) with e = .12. While this standard deviation yields a large range, it is acceptable. The range for the 210th unit was between 3,119 and 6,408 with a mean of 4,470. This range was also acceptable.

The triangle distribution is a finite distribution, Here, the highest point was .36, based on three standard deviations of the normal error term distribution used. Since the triangle distribution was to be symmetrical, the lowest point would be -.36 and the most likely point zero. This distribution resulted in all the first unit costs falling between 17,442 and 35,833 with a mean of 25,000, based on formula (18) with e = .12.

Since the Cauchy distribution has the same attributes as the normal distribution but with fatter tails, the same value, .12, was used as the scale parameter for the Cauchy error term distribution. This scale parameter resulted in approximately 80% (9:155) of the first unit costs falling within 3.2361 scales, or between 16,955 and 36,863 with a mean of 25,000, based on formula (18) with (3.2361 * 3) replacing (3 * e) and e = .12.

Cauchy Error Term Distribution Problems. After the scale parameters were selected, some trial simulation runs were made. During these simulation runs, the Cauchy error term distribution was noted to cause some very extreme values. Despite the fact that this distribution is similar to the normal distribution, it has fatter tails. The Cauchy distribution thus generated some unusual values. On one run, the unit cost ranged from 4.8×10^{-23} for unit 164 to 1.1×10^{22} for unit 44. Values became so extreme that the computer could not change the costs from logarithmic form. The logarithmic unit costs ranged from -7017.72 to 2778.005.

For the -7017.72 logarithmic value to be changed to a cost of 5 (logarithmic 1.609437912), the scale factor would have to be changed to .000117424. This scale factor would result in 80% (9:158) of the first unit costs falling within 3.2361 scales or between 24,991 and 25,010 with a mean of 25,000, based on formula (18) with (3.2361 * e) replacing (3 * e) and e = .000117424. Thus, to account for the extreme values possible with a Cauchy error term distribution, the scale factor would have to be made so small that the majority of the variation would be removed. Note that the cost of unit 44 in the previous paragraph is more than trillions of times greater than the 1987 national debt and that the cost of unit 164 in the previous paragraph is virtually zero. Given the above cited problems, the Cauchy distribution was dropped from this study and should not be used in future learning curve simulations. Cases V and VI were deleted from the thesis.

Conclusion. Parameters that reflect comparable probability bounds for the error term distributions have been selected. The normal and triangle distributions thus defined were used in the remainder of this study.

Analysis of Fitting Techniques

All techniques were analyzed based on their ability to fit and forecast simulated learning curve data. The techniques were first reviewed on the fit achieved and, second, on the forecasting performance. The review covered the following cases with the enumerated techniques in the listed order:

- Case I: equal lot sizes with normal error terms
  - ordinary least-squares technique
  - median slope technique
  - mean slope technique
- Case II: unequal lot sizes with normal error terms
  - ordinary least-squares technique
  - weighted least-squares technique
  - median slope technique
  - mean slope technique
- Case III: equal lot sizes with triangle error terms
  - ordinary least-squares technique
  - median slope technique
  - mean slope technique
- Case IV: unequal lot sizes with triangle error terms
  - ordinary least-squares technique
  - weighted least-squares technique
  - median slope technique
  - mean slope technique

Analysis of Fit. The fit of the techniques is analyzed in two areas, the parameters and the dispersion. The parameter estimates for first unit cost, a, and the learning curve coefficient, b, are compared along with the dispersion of the data points around the fitted line. These analyses are based on 100 production runs that were generated using a first unit cost, A, of 25000 and a learning curve slope, B, of -.321928095 (ln(.8)/ln(2)). These numbers are the population or true values.

Equal Lot Sizes with Normal Error Terms. Analysis of the box-and-whisker plots in Figure 4 and the data in Table 2 shows several things. First, note that for the first unit costs, the ordinary least-squares technique has the smallest range, followed by the median slope technique, with the mean slope technique having the largest range. Second, notice the distributions of the statistic about the true

Table 2. Estimated First Unit Cost for
Equal Lot Sizes with Normal Error Terms

Population or True First Unit Cost = 25,000

	Ordinary Least-Squares Technique		Median Slope Technique		Mean Slope Technique
Maximum	27,745.3	Maximum	28,434.5	Maximum	31,078.8
95%	26,004.2	95%	27,511.6	95%	29,355.0
75%	25,216.2	75%	25,545.7	75%	26,206.2
Mean	24,465.6	Median	24,864.3	Mean	25,150.3
Median	24,419.0	Mean	24,791.7	Median	25,017.5
25%	23,791.8	25%	23,569.8	25%	23,810.5
5%	22,917.5	5%	22,467.5	5%	21,977.4
Minimum	21,879.3	Minimum	21,346.4	Minimum	19,529.0
RANGE:					
Total	5,866.0		7,088.1		11,549.8
1st-3rd Quartile	1,424.4		1,975.9		2,395.7
BIAS:					
Mean	-534.4		-208.3		150.3
Median	-581.0		-135.7		17.5

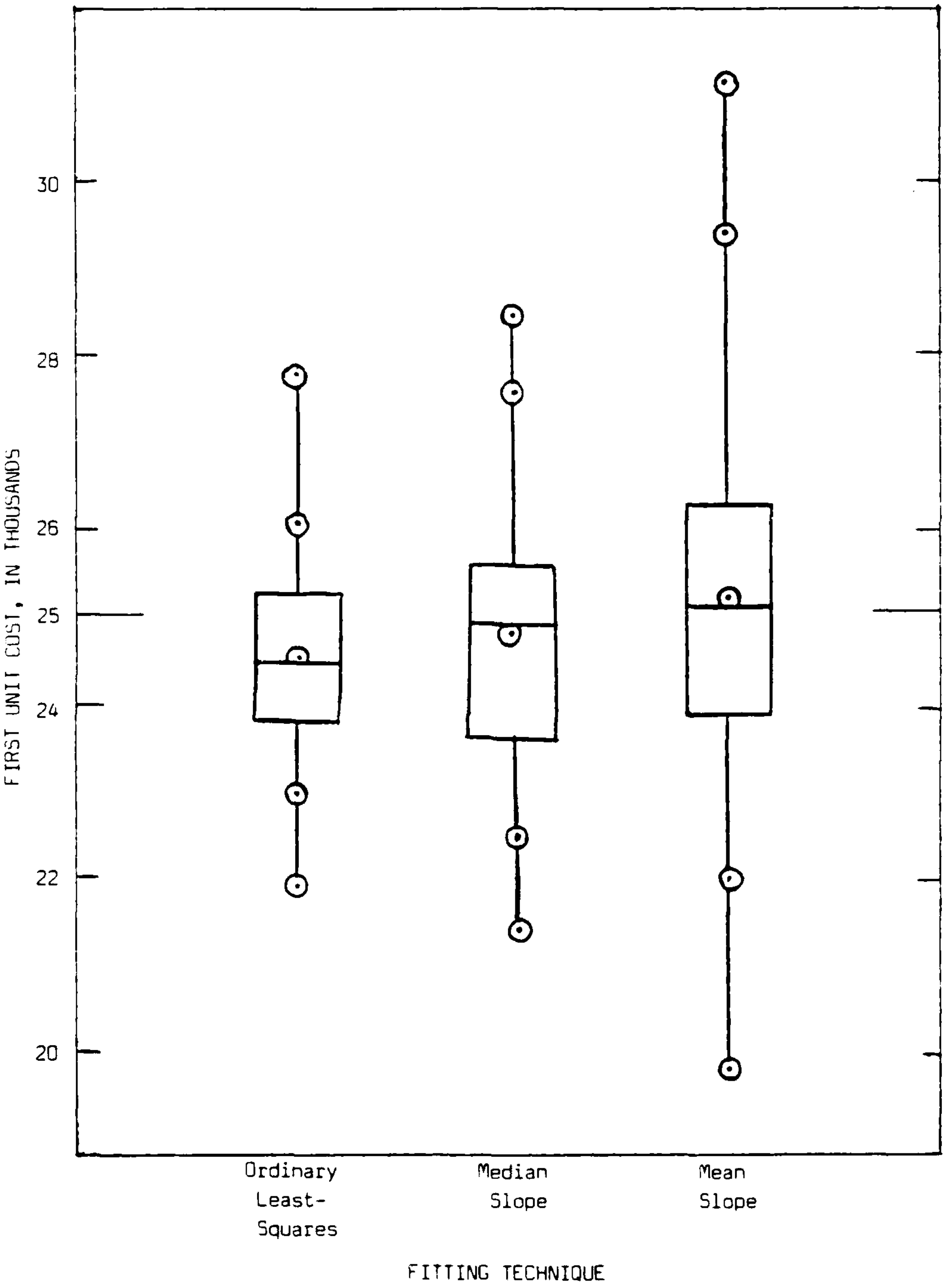


Figure 4. First Unit Cost Statistics,
Equal Lot Sizes with Normal Error Terms

value. Bias can be seen if the average and median statistics are either above or below the population value. It should be noted that, due to the nature of this research, statistical tests fot the significance of this bias were not included. Although these differences in average results from the population or true value are probably not statistically significant at a high level of confidence, the word "bias" is most descriptive. The mean slope technique has the least bias, with the average estimate only slightly high. The median slope technique has an average estimated first unit cost that is slightly low, while the ordinary least-squares technique is significantly lower.

Similar information about the statistic that estimates the learning curve coefficient is shown in Figure 5 and Table 3. Again note that the ordinary least-squares technique has the smallest range, the median slope technique has a slightly larger range and the mean slope technique has the largest range. As with the first unit cost, some techniques appear biased. The distributions for the learning curve coefficient show the mean slope technique has the least bias and is most closely centered to the population value. The median slope technique is biased slightly high and the ordinary least-squares technique is biased significantly higher. The impact of high bias on the learning curve coefficient parameter is a higher learning curve slope, i.e. less learning is estimated than the population experiences.

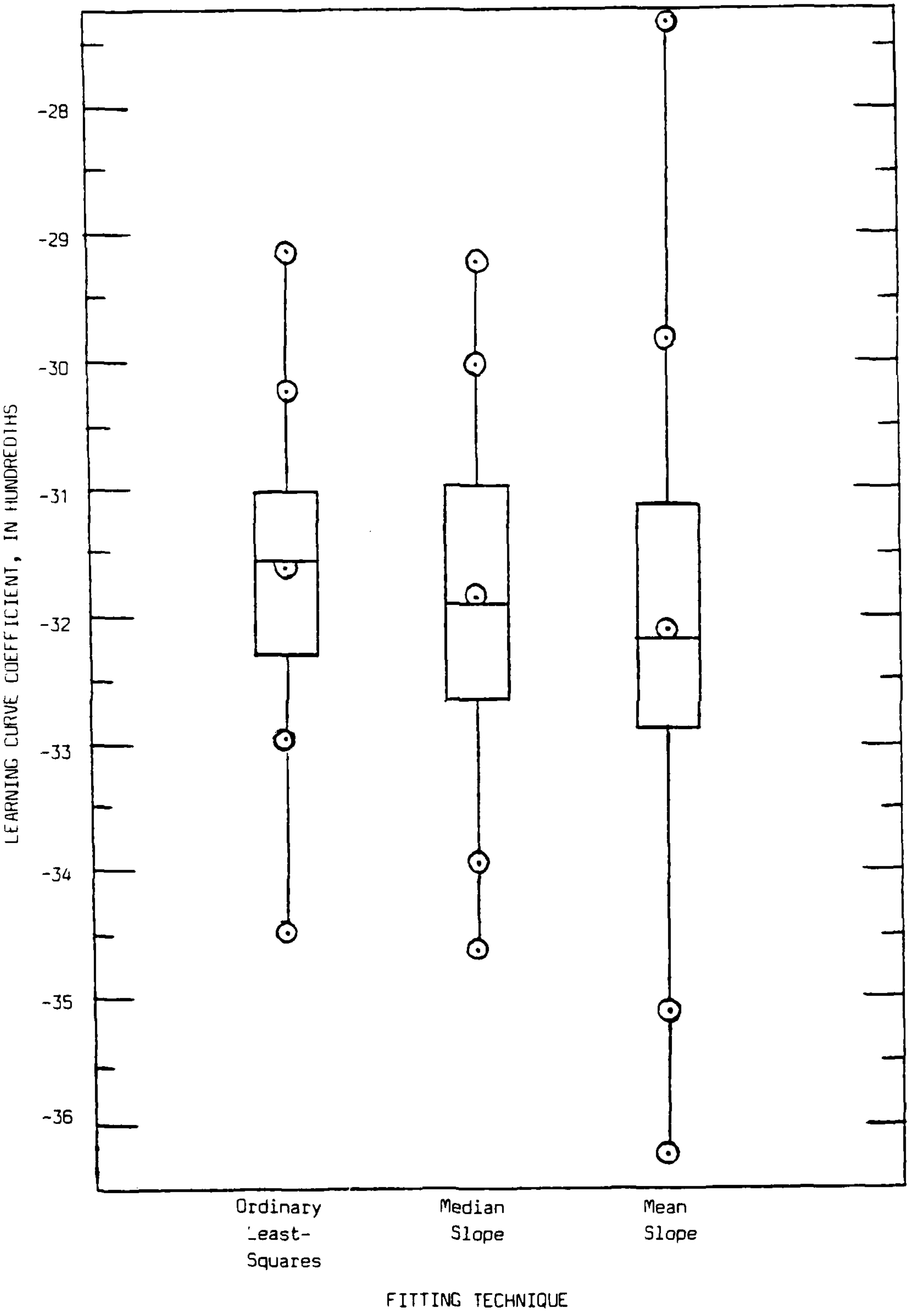


Figure 5. Learning Curve Coefficient Statistics, Equal Lot Sizes with Normal Error Terms

Table 3. Estimated Learning Curve Coefficient for Equal Lot Sizes with Normal Error Terms

Population or True Coefficient = -.321928095

	Ordinary		Median		Mean
	Least-Squares		Slope		Slope
	Technique		Technique		Technique
Maximum	-0.290998	Maximum	-0.292049	Maximum	-0.272929
95%	-0.302127	95%	-0.300358	95%	-0.297986
75%	-0.310133	75%	-0.309594	75%	-0.310900
Median	-0.315503	Mean	-0.318482	Mean	-0.321233
Mean	-0.315961	Median	-0.318849	Median	-0.321773
25%	-0.322814	25%	-0.326391	25%	-0.328686
5%	-0.329688	5%	-0.339246	5%	-0.351119
Minimum	-0.344737	Minimum	-0.346114	Minimum	-0.362477
RANGE:					
Total	0.053739		0.054065		0.089548
1st-3rd Quartile	0.012681		0.016797		0.017786
BIAS:					
Mean	0.005967		0.003446		0.000695
Median	0.006425		0.003079		0.000155

The dispersion of the data points around the learning curve line shows (see Table 4) the ordinary least-squares technique has the lowest average MAD and average MSD, the median slope technique has the next best average MAD (almost 52% larger than for ordinary least-squares) and average MSD (more than 221% larger than for ordinary least-squares) and

Table 4. Measures of Dispersion for Equal Lot Sizes with Normal Error Terms

	Ordinary	Median	Mean
	Least-Squares	Slope	Slope
	Technique	Technique	Technique
AVERAGE OF 100:			
MAD	646.7	981.1	1,143.3
MSD	102,917.0	330,897.5	525,822.9

the mean slope technique has the greatest amount of dispersion with an average MAD over 76% greater than for ordinary least-squares and an average MSD almost 411% greater than the average MSD for ordinary least-squares. The greater the MAD, the greater is the dispersion about the learning curve line. That is to say, the actual points are further above and below the fitted learning curve line. A greater MSD can be influenced by all the data points or can be more greatly influenced by one data point that is very far from the fitted line. These MAD and MSD results are not terribly surprising since ordinary least-squares attempts to best fit the line to minimize squared errors, while the other techniques do not attempt to best fit.

In conclusion, the use of the ordinary least-squares technique on Case I data resulted in the smallest range and least dispersion but the most bias in the parameters estimated. The combination of a low bias on the first unit cost and a high bias on the learning curve coefficient will yield high predictions of future costs. On the other hand, the mean slope technique generally gives the least biased estimates of the first unit cost and the learning curve coefficient, but has a much greater range and dispersion of values. On average, a less biased prediction of future costs will be made by a learning curve line fit with the mean slope technique. However, because of the larger dispersion, larger errors could occur with this technique. The median slope technique balances the problems of the other two techniques, a smaller range of possible first unit costs and learning curve coefficients than the mean slope technique with less bias than the ordinary least-squares technique.

The next three sections analyze the fitting technique results for the remaining three cases. The casual reader may wish to skip to the analysis of cost predictions (page 54) to see how the techniques predicted future costs in each case. Turn to Chapter 5 for a summary of findings for all cases.

Unequal Lot Sizes with Normal Error Terms. Analysis of the data through use of box-and-whisker plots (plots not shown, see data in Table 5) shows that the weighted least-squares technique has the smallest range for the first unit costs. The mean slope technique has the next largest range followed by the median slope technique. The ordinary least-squares technique has the largest range. In terms of bias, the techniques that were used on the equal lot size data reversed their order with the unequal lot size data. The ordinary least-squares technique now has the least bias, while the median slope technique has almost twice that amount. The mean slope technique has slightly more high bias than the median slope technique and the weighted least-squares technique has the largest high bias, about three times that of ordinary least-squares. All techniques gave high estimates for the first unit cost, i.e. they were all biased high. The impact of this high bias results in an

Table 5. Estimated First Unit Cost for
Unequal Lot Sizes with Normal Error Terms

Population or True First Unit Cost = 25,000

	Ordinary Least-Squares Technique		Median Slope Technique		Mean Slope Technique		Weighted Least-Squares Technique
Maximum	29,168.4	Maximum	28,824.8	Maximum	29,000.0	Maximum	28,825.9
95%	28,172.1	95%	27,776.9	95%	28,380.4	95%	27,900.6
75%	26,633.6	75%	26,512.7	75%	26,541.0	75%	26,680.3
Median	25,317.8	Mean	25,472.2	Median	25,577.9	Median	25,856.7
Mean	25,222.0	Median	25,432.9	Mean	25,512.0	Mean	25,787.9
25%	24,046.3	25%	24,568.9	25%	24,512.9	25%	24,930.0
5%	21,796.7	5%	22,628.7	5%	22,934.7	5%	23,700.2
Minimum	20,991.0	Minimum	21,235.0	Minimum	22,075.6	Minimum	22,122.0
RANGE:							
Total	8,177.4		7,589.8		6,924.4		6,703.9
1st-3rd Quartile	2,587.3		1,943.8		2,028.1		1,750.3
BIAS:							
Mean	222.0		472.2		512.0		787.9
Median	317.8		432.9		577.9		856.7

expected first unit cost estimate greater than the population value.

The distributions of the learning curve coefficients from the box-and-whisper plots (plots not shown, see data in Table 6) show that the mean slope technique has the smallest range followed very closely by the weighted least-squares technique. The median slope technique has a 15% range increase and the ordinary least-squares technique has over a 30% increase. The mean and median biases are split for the ordinary least-squares technique, and the values are very small. The median slope technique has low bias, three times greater than the low rating of ordinary least-squares. The

Table 6. Estimated Learning Curve Coefficient for Unequal Lot Sizes with Normal Error Terms

Population or True Coefficient = -.321928095

	Ordinary Least-Squares Technique		Median Slope Technique		Mean Slope Technique		Weighted Least-Squares Technique
Maximum	-0.279564	Maximum	-0.287639	Maximum	-0.296019	Maximum	-0.293522
95%	-0.286794	95%	-0.300991	95%	-0.305469	95%	-0.307811
75%	-0.310914	75%	-0.317822	75%	-0.317098	75%	-0.320222
Mean	-0.321115	Mean	-0.324675	Mean	-0.325050	Median	-0.326632
Median	-0.323132	Median	-0.325780	Median	-0.325549	Mean	-0.326685
25%	-0.334554	25%	-0.333160	25%	-0.333512	25%	-0.335174
5%	-0.346421	5%	-0.340082	5%	-0.342080	5%	-0.344562
Minimum	-0.353979	Minimum	-0.353451	Minimum	-0.353137	Minimum	-0.350645
RANGE:							
Total	0.074415		0.065812		0.057118		0.057123
1st-3rd Quartile	0.023640		0.015338		0.016414		0.014952
BIAS:							
Mean	0.000813		-0.002747		-0.003122		-0.004757
Median	-0.001204		-0.003852		-0.003621		-0.004704

mean slope technique has approximately the same overall low bias as the median slope technique while the weighted least-squares technique has the largest low bias, almost 45% more than for median slope. The result of low bias f)r the learning curve coefficient is a lower learning curve slope, that is the estimated decrease in cost more than the population experiences.

The dispersion of the data points (see Table 7) around the fitted learning curve line shows the ordinary least-squares technique has the lowest average MAD and average MSD. The median slope, weighted least-squares and mean slope techniques have average MADs within 60% of ordinary least-

Table 7. Measures of Dispersion for
Unequal Lot Sizes with Normal Error Terms

	Ordinary Least-Squares Technique	Median Slope Technique	Mean Slope Technique	Weighted Least-Squares Technique
AVERAGE OF 100:				
MAD	1,392.4	2,035.8	2,150.8	2,067.6
MSD	710,770.0	2,792,609.0	2,689,005.0	3,540,134.0

squares. The median slope and mean slope techniques have average MSDs over 350% greater while the weighted least-squares technique has the highest average MSD, almost five times that of ordinary least-squares. This is unexpected since weighted least-squares should better fit the data when the data has unequal weights. Resolution of this may require that the errors by lot be weighted by lot size.

In conclusion, the use of weighted least-squares on Case II data resulted in the overall least range but the most bias in the parameters estimated and the highest overall dispersion. This yielded a combination of high bias on the first unit cost and low bias on the learning curve coefficient which should on average result in low predictions of future costs. The ordinary least-squares technique has the least bias and smallest dispersion but the widest range. Tnis resulted in a combination of high bias on the first unit cost and a split, or almost no low, bias on the learning curve coefficient yielding high predictions of future costs that will become low predictions after many, many units. Both the median slope and the mean slope techniques have more bias than the ordinary least-squares technique and greater ranges over the first unit cost and learning curve coefficient than the weighted least-squares technique.

Equal Lot Sizes with Triangle Error Terms. Analysis of the box-and-whisker plots (plots not shown, see Table 8 for data) shows that the ordinary least-squares technique has the lowest range for the first unit cost. The median slope technique increases the range by almost 14% while the mean slope technique has the largest range, over 66% greater than the range of ordinary least-squares. The mean slope technique is almost centered with a very small high bias, the median slope technique has low bias and the

Table 8. Estimated First Unit Cost for Equal Lot Sizes with Triangle Error Terms

Population or True First Unit Cost = 25,000

	Ordinary Least-Squares Technique		Median Slope Technique		Mean Slope Technique
Maximum	27,943.2	Maximum	29,210.4	Maximum	31,814.6
95%	27,013.9	95%	27,783.4	95%	29,505.2
75%	25,288.1	75%	26,055.2	75%	26,503.5
Mean	24,599.4	Mean	24,844.4	Mean	25,136.7
Median	24,489.9	Median	24,727.2	Median	24,994.6
25%	23,985.1	25%	23,784.0	25%	23,411.4
5%	22,630.1	5%	22,196.5	5%	21,278.9
Minimum	20,996.2	Minimum	21,321.9	Minimum	20,134.3
RANGE:					
Total	6,947.0		7,888.5		11,680.3
1st-3rd Quartile	1,303.0		2,271.2		3,092.1
BIAS:					
Mean	-400.6		-155.6		136.7
Median	-510.1		-272.8		-5.4

ordinary least-squares technique has the most low bias, about twice the bias of the median slope technique. The impact of high bias is a higher first unit cost than the population value while the impact of low bias is a lower first unit cost than the population value. A wider range of first unit costs increases the probability of getting a first unit cost significantly different than the population value.

The distributions of the learning curve coefficients based on an analysis of the box-and-whisker plots (plots not shown, see data in Table 9) show that the ordinary least-squares technique has the smallest range. The median slope technique has a range over 12% larger and the mean slope

Table 9. Estimated Learning Curve Coefficient for Equal Lot Sizes with Triangle Error Terms

Population or True Coefficient = -.321928095

	Ordinary Least-Squares Technique		Median Slope Technique		Mean Slope Technique
Maximum	-0.285139	Maximum	-0.281546	Maximum	-0.274801
95%	-0.299257	95%	-0.297417	95%	-0.288618
75%	-0.308678	75%	-0.308188	75%	-0.307904
Mean	-0.316054	Median	-0.318006	Median	-0.319935
Median	-0.316175	Mean	-0.318215	Mean	-0.320172
25%	-0.321786	25%	-0.329008	25%	-0.334362
5%	-0.334803	5%	-0.338870	5%	-0.354159
Minimum	-0.344565	Minimum	-0.348343	Minimum	-0.359162
RANGE:					
Total	0.059426		0.066797		0.084361
1st-3rd Quartile	0.013108		0.020820		0.026458
BIAS:					
Mean	0.005874		0.003713		0.001756
Median	0.005753		0.003922		0.001993

technique has the largest range, nearly 42% greater than the range of ordinary least-squares. The mean slope technique has the smallest bias, slightly high. The median slope technique has about twice the bias and the ordinary least-squares technique has nearly triple the bias of the mean slope technique. The result of high bias for the learning curve coefficient is a higher learning curve slope and a lower rate of learning than with the population value. As with the first unit cost range, the wider the range the more often there will be a larger difference between the fitted value and the true value.

The dispersion of the data points around the fitted learning curve line (see data in Table 10) shows the ordinary least-squares technique has the lowest average MAD and average MSD. The median slope technique has an increase of 48% for the average MAD and the mean slope technique has the highest MAD, over 73% higher than for ordinary least-squares. The median slope technique has an increase of over 221% for the average MSD and the mean slope technique has the highest

Table 10. Measures of Dispersion for Equal Lot Sizes with Triangle Error Terms

	Ordinary Least-Squares Technique	Median Slope Technique	Mean Slope Technique
AVERAGE OF 100:			
MAD	847.9	1,255.3	1,468.5
MSD	162,314.7	521,723.4	827,811.2

average MSD, about 410% higher than the MSD of ordinary least-squares. As the average MAD increases, the data points are further from the fitted line. As the MSD increases, the data points are further from the fitted line, and more importance is place on data points that are far from the fitted line. Note that even when the type of error term changes, ordinary least-squares still fits the data with the smallest error.

In conclusion, the use of the ordinary least-squares technique on Case III data resulted in the smallest range and least dispersion but the most bias in the parameters estimated. The combination of low bias on the first unit cost and high bias on the learning curve coefficient will result in high predictions of future costs. On the other hand, the mean slope technique gives generally less biased estimates of the first unit cost and learning curve coefficients, but a much greater range and dispersion of values. The median slope technique has more bias than the mean slope technique and has greater range than the ordinary least-squares technique. As stated before, the result of low bias on the first unit cost and high bias on the learning curve coefficient is high predictions of future costs over the population. This bias is greatest with the ordinary least-squares technique. On average, the better predictions of future costs will be made by the mean slope technique. However, the higher dispersion and larger range allows more opportunity for the estimates to be significantly different than the population values.

Unequal Lot Sizes with Triangle Error Terms. Analysis of the box-and-whisker plots (plots not shown, see data in Table 11) shows the weighted least-squares technique has the smallest range for the first unit cost. The mean slope technique increases the range over 12% and the ordinary least-squares technique increases the range over 20%. The median slope technique has the largest range, over 34% greater than the smallest range. The ordinary least-squares technique has the least bias, with the average

Table 11. Estimated First Unit Cost for
Unequal Lot Sizes with Triangle Error Terms

Population or True First Unit Cost = 25,000

	Ordinary		Median		Mean		Weighted
	Least-Squares		Slope		Slope		Least-Squares
	Technique		Technique		Technique		Technique
Maximum	29,486.0	Maximum	29,996.0	Maximum	29,182.7	Maximum	29,287.6
95%	28,838.0	95%	28,852.6	95%	28,446.8	95%	28,419.4
75%	27,035.4	75%	27,137.3	75%	26,944.3	75%	26,801.5
Mean	25,343.4	Mean	25,568.3	Mean	25,538.3	Mean	25,812.3
Median	25,159.6	Median	25,504.3	Median	25,391.5	Median	25,687.0
25%	23,677.8	25%	24,161.3	25%	24,250.3	25%	24,752.3
5%	22,263.3	5%	22,258.6	5%	22,986.5	5%	23,552.6
Minimum	20,425.3	Minimum	19,949.7	Minimum	20,759.7	Minimum	21,793.3
RANGE:							
Total	9,060.7		10,046.3		8,423.0		7,494.3
1st-3rd Quartile	3,357.6		2,976.0		2,694.0		2,049.2
BIAS:							
Mean	343.4		568.3		538.3		812.3
Median	159.6		504.3		391.5		687.0

estimated first unit cost somewhat above the population value. The mean slope technique has higher bias, about double, and the median slope technique has even higher bias, a little more than double. The weighted least squares technique has the largest high bias, about three times greater than for ordinary least-squares. The result of high bias is higher first unit costs than the population value.

The dispersions of the learning curve coefficients from the box-and-whisker plots (plots not shown, see data in Table 12) show that the weighted least-squares technique has the lowest range. The mean slope technique has a higher range while the median slope technique has the next to the largest

Table 12. Estimated Learning Curve Coefficient for Unequal Lot Sizes with Triangle Error Terms

Population or True Coefficient = -.321928095

	Ordinary Least-Squares Technique		Median Slope Technique		Mean Slope Technique		Weighted Least-Squares Technique
Maximum	-0.277082	Maximum	-0.277520	Maximum	-0.286985	Maximum	-0.292247
95%	-0.292436	95%	-0.298026	95%	-0.304547	95%	-0.305739
75%	-0.307230	75%	-0.313693	75%	-0.315566	75%	-0.316435
Mean	-0.321065	Mean	-0.324275	Median	-0.323399	Median	-0.324918
Median	-0.321460	Median	-0.324411	Mean	-0.324171	Mean	-0.325823
25%	-0.336550	25%	-0.335525	25%	-0.334843	25%	-0.336008
5%	-0.349381	5%	-0.348340	5%	-0.345254	5%	-0.347735
Minimum	-0.356997	Minimum	-0.353926	Minimum	-0.354738	Minimum	-0.352516
RANGE:							
Total	0.079915		0.076406		0.067753		0.060269
1st-3rd Quartile	0.029320		0.021832		0.019277		0.019573
BIAS:							
Mean	0.000863		-0.002347		-0.002243		-0.003895
Median	0.000468		-0.002483		-0.001471		-0.002990

range. The ordinary least-squares technique has the highest range, about 32.6% greater than for weighted least-squares. However, the ordinary least-squares technique has the least bias, with the average estimated learning curve coefficient only slightly above the population or true value. The other techniques are biased low, or below the population value. The mean slope technique has the smallest low bias, about three times that of ordinary least-squares. The median slope technique has essentially the same low bias as the mean slope technique and the weighted least-squares technique has the most low bias, about six times that of ordinary least-squares. The result of low bias on the learning curve coefficient is a lower learning curve slope than for the population, the estimated decrease in costs is thus more than the population experiences. The result of high bias on the learning curve coefficient is a higher learning curve slope than that of the population, the estimated decrease in costs is less than the population experiences.

The dispersion of the data points around the fitted learning curve line shows (see Table 13) the ordinary least-squares technique fits best on average using MAD and MSD. The median slope technique has an average MAD increase over 43%, the weighted least-squares technique has an average MAD increase of almost 49% and the mean slope technique has the highest MAD, an average MAD increase over 56% that of ordinary least-squares. The mean slope technique has an

Table 13. Measures of Dispersion for
Unequal Lot Sizes with Triangle Error Terms

	Ordinary Least-Squares Technique	Median Slope Technique	Mean Slope Technique	Weighted Least-Squares Technique
AVERAGE OF 100:				
MAD	1,571.1	2,252.2	2,454.1	2,336.2
MSD	816,844.6	3,125,297.0	3,085,647.0	4,163,786.0

average MSD almost 278% higher, the median slope technique has an average MSD nearly 283% higher and the weighted least-squares technique has the highest average MSD, nearly 410% higher than for ordinary least-squares. When comparing increases, the larger increase of the average MAD of the mean slope technique versus the lower increase of the average MSD shows that the data points are all equally dispersed, with fewer outlying points. The greater the dispersion of data points around the fitted learning curve line, the greater the possibility of the future costs of units being far from the fitted line value.

In conclusion, the use of the weighted least-squares technique on Case IV data gives the smallest range and the second smallest dispersion but the most bias in the parameters estimated. This technique yielded a combination of high bias on the first unit cost and low bias on the learning curve coefficient which would result in low predictions of future costs. The ordinary least-squares technique generally gives the least biased estimates of the first unit cost and the learning curve coefficient and the least dispersion, but a much larger range of parameter values. This situation yielded a combination of high bias on the first unit cost and high bias on the learning curve coefficient which results in high predictions of future costs. The median slope and mean slope techniques have more bias than the ordinary least-squares technique, wider ranges than the weighted least-squares technique and more dispersion than both the least-squares techniques. The predictions from the median slope and mean slope techniques are in between the weighted and ordinary least-squares techniques results.

Analysis of Cost Predictions. The various learning curve fitting techniques were compared to see how each technique predicted future costs. The data used to do these comparisons were the predictions from the 100 production runs. These predictions were also compared to the population or true future unit cost. No error term was included in these true future costs since the true cost is known. Independence of the error terms assures that the comparison of predictions to the true population value is sufficient. The cases are discussed in the same order as in the prior section.

Equal Lot Sizes with Normal Error Terms. Analysis of predicted future costs was done on the predictions of costs for unit 225 and unit 800.

The analysis showed that the average predicted costs of unit 225 (see Table 14.) are above the population cost for

Table 14. Predicted Cost of Unit 225 for Equal Lot Data with Normal Error Terms

	Mean Prediction	Range
Ordinary Least-Squares	4415.10	343.34
Median Slope	4410.38	570.24
Mean Slope	4401.54	578.50
Population Cost	4372.26	

all techniques, but less than 1% above. The mean slope technique's average prediction is closest to the population cost. The median slope technique's average prediction is a little higher than the prediction from the mean slope technique. The ordinary least-squares technique's average prediction is the highest and furthest from the population cost. On the other hand, the ordinary least-squares technique has the smallest range of predictions, the median slope technique has a range of predictions over 66% larger and the mean slope technique has the largest range of predictions (over 68% above the ordinary least-squares range).

The analysis of the costs of unit 800 (see Table 15.)

Table 15. Predicted Cost of Unit 300 for Equal Lot Data with Normal Error Terms

	Mean Prediction	Range
Ordinary Least-Squares	2958.32	400.30
Median Slope	2945.48	465.18
Mean Slope	2929.32	503.24
Population Cost	2906.39	

shows that the average predicted costs were within 2% above the population cost of unit 800 for all techniques. The mean slope technique's average prediction is closest to the population cost, about .79% higher. The median slope technique's average prediction is the next higher, about 1.34% greater than the population cost. The ordinary least-squares technique's average prediction is the highest, almost 1.79% above the population cost. On the other hand, the ordinary least-squares technique again has the smallest range of predicted cost, the median slope technique has a range over 16% higher and the mean slope techniques has the highest range, almost 25% higher.

The predictions of future costs for units 225 and 800 in Case I are consistent with the prior analysis of the fit of the three techniques. The ordinary least-squares technique's average prediction of cost is getting further from the population cost at a faster rate than for the median and mean slope technique. The ranges of the predicted values for the different techniques more similar.

Unequal Lot Sizes with Normal Error Terms. Analysis of predicted future costs was done on the predictions of costs for unit 225 and unit 800.

The analysis shows that the average predicted costs of unit 225 (see Table 16.) are above the population cost for all techniques, but less than 1.1% above. The mean slope technique's average prediction is closest to the population

Table 16. Predicted Cost of Unit 225 for Unequal Lot Data with Normal Error Terms

	Mean Prediction	Range
Ordinary Least-Squares	4419.45	499.10
Median Slope	4382.93	591.42
Mean Slope	4380.43	482.00
Weighted Least-Squares	4390.09	304.74
Population Cost	4372.26	

cost while the median slope technique's average prediction is a little higher. The weighted least-squares technique's average prediction is more than double the bias of the mean slope technique's prediction. On the other hand, the ordinary least-squares technique predicted a mean cost that was the highest and furthest from the population cost, almost six times the bias of the mean slope technique. The weighted least-squares technique has the smallest range of predictions, the mean slope technique has a range over 58% larger, the ordinary least-squares technique has a range 64% larger and the median slope technique has the largest range, over 94% larger than the range of weighted least-squares.

The analysis of the costs of unit 800 (see Table 17.) shows that the average predicted costs are within 1.3% above and .2% below the population cost of unit 800. The median slope technique's average prediction is closest to the population cost, about .07% lower. The weighted least-squares technique's average prediction is the next lower, more than double the median slope difference. The mean slope

Table 17. Predicted Cost of Unit 800 for Unequal Lot Data with Normal Error Terms

	Mean Prediction	Range
Ordinary Least-Squares	2942.63	607.59
Median Slope	2904.16	517.34
Mean Slope	2900.89	427.36
Weighted Least-Squares	2901.38	380.43
Population Cost	2906.39	

technique's average prediction is lowest, just .19% below the population cost. The ordinary least-squares technique's average prediction is above the population value, over 1.24% higher. On the other hand, the weighted least-squares technique has the smallest range of predicted cost, the mean slope technique has a range over 12% higher, the median slope technique has a range about 36% higher and the ordinary least-squares technique has a range almost 60% higher than the range of the weighted least-squares technique.

The predictions of future costs for units 225 and 800 in Case II are consistent with the prior analysis of the fit of the four techniques. Because of the combinations of bias, range and distribution, the prior ranking of techniques is not followed. These combinations result in the best prediction from a technique that has mid-level ratings for bias, range and dispersion. The weighted least-squares technique, which had the most bias and smaller ranges did not have the highest predictions. The median slope, weighted least-squares and mean slope techniques went from predicting above the population to predicting costs below the population cost. These techniques will predict costs further below the population for units greater than unit 800. The ordinary least-squares technique has mean predicted costs that are increasingly above the population value as the unit number gets larger.

Equal Lot Sizes with Triangle Error Terms. Analysis of predicted future costs was done on the predictions of costs for unit 225 and unit 800.

The analysis shows that the average predicted costs of unit 225 (see Table 18.) are above the population cost for all techniques, but less than 1.5% above. The mean slope technique's average prediction is closest to the population cost. The median slope technique's average prediction is a little higher than for the mean slope technique. The ordinary least-squares technique's average prediction is highest and furthest from the population cost, almost 1.5% above the population cost. On the other hand, the ordinary least-squares technique has the smallest range of

Table 18. Predicted Cost of Unit 225 for Equal Lot Data with Triangle Error Terms

	Mean Prediction	Range
Ordinary Least-Squares	4436.37	345.49
Median Slope	4425.54	564.68
Mean Slope	4418.77	588.37
Population Cost	4372.26	

predictions, the median slope technique has a range over 63% higher and the mean slope technique has the largest range of predicted costs, over 70% above ordinary least-squares.

The analysis of the costs of unit 800 (see Table 19.)

Table 19. Predicted Cost of Unit 800 for Equal Lot Data with Triangle Error Terms

	Mean Prediction	Range
Ordinary Least-Squares	2971.77	374.78
Median Slope	2956.74	537.84
Mean Slope	2945.43	608.65
Population Cost	2906.39	

shows that the average predicted costs are within 2.3% above the population cost of unit 800 for all techniques. The mean slope technique's average prediction is closest to the population cost, over 1.3% greater. The median slope technique's average prediction is higher, over 1.7% greater than the population cost. The ordinary least-squares technique's average prediction is the highest, over 2.2% above the population cost. On the other hand, the ordinary least-squares technique has the smallest range of predicted cost, the median slope technique has a range over 43% higher and the mean slope technique has the largest range, over 62% higher than the range of ordinary least-squares.

The predictions of future costs for unit 225 and 800 in Case III are consistent with the prior analysis of the fit of the three techniques. The ordinary least-squares technique mean predicted cost is getting further from the population cost at a faster rate than for the median and mean slope technique. The ranges of the predicted values for the different techniques are converging.

Unequal Lot Sizes with Triangle Error Terms. Analysis of predicted future costs was done on the predictions of costs for unit 225 and unit 800.

The analysis shows that the average predicted costs of unit 225 (see Table 20.) are above the population cost for all techniques, but less than 1.6% above. The mean slope technique's average prediction is closest to the population cost. The median slope technique's average prediction is slightly higher while the weighted least-squares technique's average prediction is a little above the prediction from the mean slope technique. On the other hand, the ordinary least-squares technique has a mean predicted cost that is the highest and furthest from the population cost, almost double

Table 20. Predicted Cost of Unit 225 for Unequal Lot Data with Triangle Error Terms

	Mean Prediction	Range
Ordinary Least-Squares	4439.13	443.57
Median Slope	4403.83	721.50
Mean Slope	4403.71	655.01
Weighted Least-Squares	4413.36	324.83
Population Cost	4372.26	

the mean slope deviation. The weighted least-squares technique has the smallest range of predictions, the ordinary least-squares has a range over 37% higher, the mean slope technique has a range almost 102% higher and the median slope technique has the highest range, over 122% above that of weighted least-squares.

The analysis of the costs of unit 800 (see Table 21.) shows that the average predicted costs are within 1.8% above the population cost of unit 800 for all techniques. The mean slope technique's average prediction is closest to the population cost, over .45% higher. The mean slope technique's average prediction is slightly higher while the weighted least-squares technique's average prediction is the next higher, over .47% above the population cost. The ordinary least-squares technique's average prediction is the highest above the population cost, over 1.71% higher. On the other hand, the weighted least-squares technique has the smallest range of predicted cost, the mean slope technique

Table 21. Predicted Cost of Unit 300 for Unequal Lot Data with Triangle Error Terms

	Mean Prediction	Range
Ordinary Least-Squares	2956.15	584.82
Median Slope	2919.72	642.79
Mean Slope	2919.60	530.64
Weighted Least-Squares	2920.19	398.58
Population Cost	2906.39	

has a range about 33% higher, the ordinary least-squares technique has a range over 46% higher and the median slope technique has the largest range, over 61% higher than the range of weighted least-squares.

The predictions of future costs for units 225 and 800 in Case IV are consistent with the prior analysis of the fit of the four techniques. However, as with predictions in Case II, the data follows the general fit but the rank order is not based on the rank of the biases and ranges. The best predictions were from a technique with mid-level bias, range and dispersion ratings. The median slope, weighted least-squares and mean slope techniques are going from predicting above the population to predicting costs below the population cost. These techniques will go below and get further below the population cost for predictions of costs beyond unit 800. The ordinary least-squares technique predicts costs that are getting more above the population value and the predicted costs will be further above the population value for units beyond 800.

V. Conclusions and Recommendations

Conclusions

The analysis of the fitting techniques for the four cases yields no clear cut best technique. There is no best technique overall and no best technique in any of the four cases reviewed. As shown in Tables 22 and 23, there are tradeoffs for each situation.

As can be seen in Table 22, when data is from Case I, equal lot sizes with normal error terms, the ordinary least-squares technique has the most bias but the smallest range while the mean slope technique has the least bias but the largest range. The choice of a technique for prediction must recognize these tradeoffs. The mean slope technique, on average, will provide a better estimate of future costs but with a greater chance of larger deviations. On the other hand, the ordinary least-squares technique will provide less chance of large deviations but predictions will be higher than population values on average.

As can be seen in Table 22, when data is from Case II, unequal lot sizes with normal error terms, the situation is more complex. The weighted least-squares technique has the most biased estimates, yet the tightest ranges, for the parameters; it also has the tightest prediction ranges and good average predictions. The average prediction gets better in comparison to the other techniques when the unit to be

Table 23. Summary of Findings for Triangle Error Terms

Case	Technique	A	B	Predicted Cost Unit 225	Predicted Cost Unit 800
Equal Lot Sizes	Ordinary Least-Squares	Most Low Bias and Tightest Range	Most High Bias and Tightest Range	Biased High 1.5% and Tightest Range	Biased High 2.2% and Tightest Range
	Median Slope	Less Low Bias but Wider Range	Less High Bias but Wider Range	Biased High 1.2% and 63% More Range	Biased High 1.7% and 43% More Range
	Mean Slope	Bias Split- see Table 8, Widest Range	Smallest High Bias and Widest Range	Biased High 1.1% and 70% More Range	Biased High 1.3% and 62% More Range
Unequal Lot Sizes	Ordinary Least-Squares	Smallest High Bias, but Wider Range	High Bias and Widest Range	Biased High 1.5% and 37% More Range	Biased High 1.7% and 46% More Range
	Median Slope	Biased Higher and Widest Range	Biased Low, but Wider Range	Biased High .7% and 122% More Range	Biased High .46% and 61% More Range
	Mean Slope	Biased Higher, but Wider Range	Smallest Low Bias, but Wider Range	Biased High .7% and 102% More Range	Biased High .45% and 33% More Range
	Weighted Least-Squares	Most High Bias and Tightest Range	Most Low Bias and Tightest Range	Biased High .9% and Tightest Range	Biased High .47% and Tightest Range

Percentages are approximate

predicted is further from the data. For units close to the data the average mean slope and median slope technique predictions were closest to the population or true cost. As the unit being predicted gets further from the data, the median slope technique has the best average prediction. These three techniques go from predicting above the population cost to predicting below the population cost as the unit being predicted is further from the data. The ordinary least-squares technique has the highest average prediction when close to the data and average predictions from this technique get higher above the population cost when the unit being predicted is further from the data. The choice of a technique for future predictions should recognize these tradeoffs. The mean slope technique, on average, will provide a less biased prediction of future costs close to the data but with a greater chance of larger deviations. The median slope technique, on average, will provide a less biased prediction of future costs but with the greatest chance of larger deviations, when units are close to the data. These techniques have average predictions that go from above to below the population cost. On the other hand, the weighted least-squares technique will provide less chance of larger deviations but the predictions will be a little further from the population values on average.

As can be seen in Table 23, when data is from Case III, equal lot sizes with triangle error terms, the ordinary

Table 22. Summary of Findings for Normal Error Terms

Case	Technique	A	B	Predicted Cost Unit 225	Predicted Cost Unit 800
Equal Lot Sizes	Ordinary Least-Squares	Most Low Bias, but Tightest Range	Most High Bias, but Tightest Range	Biased High 1% and Tightest Range	Biased High 1.8% and Tightest Range
	Median Slope	Less Low Bias, Wider Ranger	Less High Bias, Slightly Larger Range	Biased High .9% and 66% More Range	Biased High 1.3% and 16% More Range
	Mean Slope	Smallest Bias but High, Widest Range	Smallest High Bias, Widest Range	Biased High .7% and 68% Widest Range	Biased High .8% and 25% Widest Range
Unequal Lot Sizes	Ordinary Least-Squares	Smallest High Bias, but Widest Range	Bias Split-see Table 2, Widest Range	Biased High 1.1% and 64% More Range	Biased High 1.2% and 60% More Range
	Median Slope	Bias Higher, Range Smaller	Low Bias, But Smaller Range	Biased High .2% and 94% More Range	Biased Low .08% and 36% More Range
	Mean Slope	Bias Higher, Range Smaller	Bias Lower, Smallest Range	Biased High .19% and 58% More Range	Biased Low .2% and 12% More Range
	Weighted Least-Squares	Most High Bias, but Tightest Range	Most Low Bias, almost Smallest Range	Biased High .4% and Tightest Range	Biased Low .17% and Tightest Range

Percentages are approximate

least-squares technique has the most bias, tightest ranges and least dispersion. This yielded the tightest ranges and the most biased average predictions. The mean slope technique has the least bias, widest ranges and most dispersion. This yielded the widest range and the least biased average predictions. The median slope technique was between the other two techniques. The choice of a technique to base future predictions on must recognize these problems. The mean slope technique, on average, will provide a better estimate of future costs but with a greater chance of large deviations. On the other hand, the ordinary least-squares technique will provide less chance of large deviations but the predictions will be higher than population values on average.

As can be seen in Table 22, when data is from Case IV, equal lot sizes with triangle error terms, the situation is again more complex. The weighted least-squares technique has the tightest ranges, the most biased estimates of the parameters and close to the most dispersion, or worst fit. However, weighted least-squares has the tightest prediction ranges and good average predictions. The average prediction gets closer to the population cost as the unit being predicted is further from the data. The average predictions were closest to the population cost for the mean slope and the median slope techniques. These three techniques are predicting average costs closer to the population cost as the unit is further from the data. These average costs will go below the population cost somewhere beyond unit 800. The ordinary least-squares technique has the highest average prediction when close to the data. Average predictions from this technique get higher above the population cost with larger unit numbers. The choice of a technique must recognize these tradeoffs. The mean slope and the median slope techniques, on average, will provide good predictions of future costs but each technique has a greater chance of larger deviations. These techniques have average predictions that go from above to closer above the population cost and will go below the population costs somewhere beyond unit 800. On the other hand, the weighted least-squares technique will provide less chance of larger deviations but the predictions will be a little further from the population values on average.

At this point, the analyst faced with a learning curve problem should feel comfortable that either the ordinary least-squares or the weighted least-squares technique can provide reasonably good estimates of the learning curve equation parameters (A and B). A quick look at the relative significance indicates that the bias in the estimate of the first unit cost (A) is more significant than the bias in the estimate of the learning curve coefficient (B). The mean slope and the median slope techniques offer promise of less bias of the first unit cost and learning curve coefficient estimates, at the expense of increased dispersion in the estimates. Since the mean slope and median slope fitting techniques are not included in any learning curve programs currently available to the analysts, the application of these techniques must be postponed.

In summary, when data is in the most common form, that of unequal lot sizes, the weighted least-squares technique does provide good predictions with a tight range. The ordinary least-squares technique has a larger bias that increases with the distance from the data, but still provides good predictions.

Recommendations for Further Research

There are several areas that can use further study. First, the difference between the actual lot plot point and the heuristic lot plot point for the first lot should be investigated. How close does the heuristic approximate the unit that has the average cost of the first lot. The difference may be only slight but the impact has not been addressed. If the heuristic is not a good approximator, then new analysis should be done to see if an improved algorithm would improve the performance of the fitting techniques. Second, the weighted least-squares technique should be investigated further. Analysis should concentrate on finding a more accurate weighting of each lot. A more accurate weighting should theoretically result in a better fit of unequal lot data. This possibility, too, should be checked.

Third, additional fitting techniques should be investigated. The additional techniques should include nonlinear regression and other techniques not identified by this thesis. These new techniques should be compared among themselves and to the techniques investigated in this thesis. Fourth, statistical testing for significance of the bias should be considered. Is a bias of two percent statistically significant? Is this apparent bias due to some other factor? The answer to these questions depends on the experimental design and sample size. Resolution of this question might require replication and expansion of the study from 100 production runs to 500 or 1000. Fifth, a program should be written so that the mean slope and the median slope techniques can be applied to real data. The comparison of results from these two techniques and the ordinary least-squares and the weighted least-squares techniques could be useful.

Bibliography

1. Air Force Institute of Technology. Learning Curve Analysis. A text on cost improvement curves for QMT345. Wright-Patterson AFB OH: School of Systems and Logistics, August 1976.

2. Air Force Systems Command. The AFSC Cost Estimating Handbook Series: Volume I "AFSC Cost Estimating Handbook". Reading MA: The Analytic Sciences Corporation, undated.

3. Bryan, Noreen S. and Dr. Rolf Clark. "Is Cost Growth Being Reinforced?," Concepts, 4 (2): 105-118 (Spring 1981).

4. Conover, W. J. Practical Nonparametric Statistics (Second Edition). New York: John Wiley & Sons, 1930.

5. -----. Practical Nonparametric Statistics. New York: John Wiley & Sons Inc., 1971.

6. Gansler, Jacques S. "Defense Program Instability: Causes, Costs, and Cures," Defense Management Journal, 22 (2): 3-11 (Second Quarter 1986).

7. Hayes, Robert H. and Steven C. Wheelwright. Restoring Our Competitive Edge: Competing Through Manufacturing. New York: John Wiley & Sons, 1984.

8. Hollander, Myles and Douglas A. Wolfe. Nonparametric Statistical Methods. New York: John Wiley & Sons, 1973.

9. Johnson, Norman L. and Samuel Kotz. Continuous Univariate Distributions-1. Boston: Houghton Mifflin Company, 1970.

10. Johnston, J. Econometric Methods. New York: McGraw-Hill Book Company, 1984.

11. Kankey, Roland D. "Learning Curves: An Overview," National Estimator, 4 (2): 16-19 (Spring 1983).

12. Kankey, Roland D. and Patrick A. Thompson. Loss Functions and Forecast Accuracy Statistics: Some Measures Used to Compare Forecasting Techniques. Working Paper Series 86-91. Columbus OH: College of Administrative Science, The Ohio State University, September 1986.

13. Moskowitz, Herbert and Gordon P. Wright. Statistics for Management and Economics. Columbus OH: Charles E. Merril Company, 1935.

14. Murphy, Richard L., Assistant Professor of Quantitative Management Techniques. Personal Interview. School of Systems and Logistics, Wright-Patterson AFB OH, 9 December 1986.

15. Neter, John and others. Applied Linear Regression Models. Homewood IL: Irwin, 1983.

16. SAS User's Guide: Basics, Version 5 Edition. SAS Institute Inc., Cary NC, 1985.

17. SAS User's Guide: Statistics, Version 5 Edition. SAS Institute Inc., Cary NC, 1985.

18. Tukey, John W. Exploratory Data Analysis. Reading MA: Addison-Wesley Publishing Company, 1977.

19. Wright, T. P. "Factors Affecting the Cost of Airplanes." Journal of the Aeronautical Sciences, 3 (4): 122-128 (February 1936).

Vita

Captain Charles R. Avinger was born on 31 July 1956 in Warren, Ohio. He graduated from high school in Panama City, Florida, in June 1974. He was awarded a four-year Air Force ROTC scholarship and attended Bradley University for two years. He graduated from Colorado State University in May 1978 with a Bachelor of Science in Business Administration with a specialization in accounting. Upon graduation he received his commission in the USAF. After he was called to active duty in October 1978, he was employed as a major defense system cost analyst at HQ Electronic Systems Division. He then served as an audit team leader at Detachment 517, Kirtland AFB, New Mexico. His auditing career continued at Detachment 440, Nellis AFB, Nevada, until entering the School of Systems and Logistics, Air Force Institute of Technology, in June 1986.

Permanent address: 7322 Ross Drive
Colorado Springs, Colorado 80920

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

A187 680

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION
UNCLASSIFIED

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3. DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release;
distribution unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
AFIT/GSM/LSQ/87S-3

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
School of Systems
and Logistics

6b. OFFICE SYMBOL *(If applicable)*
AFIT/LSQ

7a. NAME OF MONITORING ORGANIZATION

6c. ADDRESS *(City, State, and ZIP Code)*
Air Force Institute of Technology (AU)
Wright-Patterson AFB, Ohio 45433-6583

7b. ADDRESS *(City, State, and ZIP Code)*

8a. NAME OF FUNDING / SPONSORING ORGANIZATION

8b. OFFICE SYMBOL *(If applicable)*

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c. ADDRESS *(City, State, and ZIP Code)*

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO	WORK UNIT ACCESSION NO.

11. TITLE *(Include Security Classification)*
ANALYSIS OF LEARNING CURVE FITTING TECHNIQUES

12. PERSONAL AUTHOR(S)
Charles R. Avinger, Captain USAF

13a. TYPE OF REPORT
MS Thesis

13b. TIME COVERED
FROM ________ TO ______

14. DATE OF REPORT *(Year, Month, Day)*
1987 September

15. PAGE COUNT
87

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP
05	03	

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
Cost Estimates, Learning Curves, Least Squares Method, Curve Fitting, Nonparametric Analysis

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

Thesis Advisor: Roland D. Kankey
Assistant Professor of Quantitative Management

Approved for public release: IAW AFR 190-1.
Lynn Wolaver
LYNN E. WOLAVER 24 Sept 87
Dean for Research and Professional Development
Air Force Institute of Technology (AFC)
Wright-Patterson AFB OH 45433

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
UNCLASSIFIED

22a. NAME OF RESPONSIBLE INDIVIDUAL
Roland D. Kankey

22b. TELEPHONE *(Include Area Code)*
(513) 878-8256

22c. OFFICE SYMBOL
AFIT/LSQ

Previous editions are obsolete.

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

BLOCK 19

ABSTRACT

This thesis was basic research on learning curve fitting techniques. The unit formulation of the learning curve was fit using two parametric techniques, ordinary least-squares and weighted least-square, and two nonparametric techniques, called median slope and mean slope. Comparisons were made between the techniques in four data cases, equal and unequal lot sizes with normally distributed error terms and equal and unequal lot sizes with triangularly distributed error terms. The Cauchy error term distribution was tried but rejected.

The fitting techniques were compared on their estimation of the formula parameters, the dispersion of the data points around the fitted line and their predictions of future costs.

Analysis showed that in cases of equal lot sizes, ordinary least-squares has the most bias, but still is a good predictor of future costs. In cases of unequal lot sizes the weighted least-squares and the ordinary least-squares techniques performed well. Ordinary least-squares estimated the parameters with the least bias but had prediction errors that increased with the unit numbers. Weighted least-squares estimated the parameters with the most bias and had prediction errors that, on average, started high and turned low as the unit number got larger. The nonparametric techniques did a better job of predicting future costs, in that the mean predictions were closer to the population costs. However, these techniques had wider ranges of predictions.

Because this was basic research, there are many areas for further research.